Adversarial Machine Learning Attack on Machine Learning-based Controller for Solar Inverters

Joaquin Massa
Dept. of Electrical Engineering and Computer Science
Texas A&M University-Kingsville
Kingsville, USA
joaquin.massa@students.tamuk.edu

Kyoung-Tak Kim
Dept.. of Electrical Engineering
Soongsil University
Seoul, Republic of Korea
whitechii@soongsil.ac.kr

Taesic Kim*
Dept. of Electrical Engineering and Computer Science
University of Missour-Columbia
Columbia, USA
tkx96@missouri.edu

BoHyun Ahn
Dept. of Electrical Engineering and Computer Science
Texas A&M University-Kingsville
Kingsville, USA
bohyun.ahn@students.tamuk.edu

Kumar Venayagamoorthy
Dept. of Electrical and Computer Engineering
Clemson University
Clemson, USA
gvenaya@clemson.edu

Jianwu Zeng
Dept. of Electrical and Computer Engineering and Technology
Minnesota State University, Mankato
Mankato, USA
jianwu.zeng@mnsu.edu

2024 IEEE Energy Conversion Congress and Exposition (ECCE) | 979-8-3503-7606-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ECCE55643.2024.10861195

***Abstract*—Adversarial machine learning (ML) attacks are stealthy attacks designed to mislead the ML model results. This paper explores adversarial ML attacks that generate adversarial noisy input data in an ML-based controller in a solar inverter. Three types of ML models, long short-term memory (LSTM), gated recurrent unit (GRU)), and bidirectional-LSTM (Bi-LSTM), are designed to replace proportional-integral (PI) controller-based vector control for a solar inverter and two white-box adversarial ML attacks (Basic Iterative Method (BIM) attack and Fast Sign-Gradient Method (FGSM)) are applied to the ML controllers. It is observed that the adversary ML attacks designed in stealthy way do not affect the PI controller, while significantly degrading performance of the ML-based controllers. Moreover, the BIM attack is more effective than FGSM and Bi-LSTM-based controller is relatively robust to the attacks compared to peer.**

Keywords—adversarial machine learning attack, machine learning, cybersecurity, solar inverter

I. Introduction

Recently, machine learning (ML)-based controllers have been proposed to replace conventional controllers in power electronics (PE) by improving the controller performance [1]-[4]. In [3], ML-based model predictive controls (MPC) are used for a three-phase inverter with L-C filter to reduce the computation of the MPC. By adopting ML accelerating hardware such as Field Programmable Gate Array (FPGA) in the PE controller board, it is expected that more and more ML-based controllers will be adopted in PE [4]. Meanwhile, the ML trustworthiness has been threatened by adversarial ML (AML) attacks that intentionally mislead the ML model results by generating adversarial data which are stealthier not to be detected by a bad data detection or an intrusion detection system [5]. Furthermore, significant cybersecurity concerns have arisen in the networked PE devices due to the extensive information exchange and firmware updates [6]. It is demonstrated that the controller input data can be spoofed to degrade the inverter operation by an adversary through the controller firmware modification [7]. It is anticipated that an adversary who can succeed in a malicious controller firmware modification can generate a stealthy controller input data spoofing attack targeting the ML-based controller. However, the impact of the AML attacks targeting ML-based PE controllers has been less studied.

This paper investigates the impact of AML attacks on an ML-based controller within a solar inverter, serving as a case study for solar inverters. Specifically, three ML models, 1) Long Short-Term Memory (LSTM [8]), 2) Gated Recurrent Unit (GRU), and 3) Bidirectional LSTM (Bi-LSTM) are tested to replace the conventional proportional-integral (PI) controller used in vector control for a solar inverter. Two white-box AML attacks, namely the Basic Iterative Method (BIM) and the Fast Gradient Sign Method (FGSM), are employed to evaluate the vulnerability of these ML-based controllers. The results indicate that while the stealthily crafted adversarial attacks do not impact the PI controller, they significantly impair the performance of the ML-based controllers. Furthermore, the BIM attack proves to be more effective than the FGSM, and among the ML-based controllers, the Bi-LSTM model exhibits relatively higher robustness against these adversarial attacks compared to the GRU and LSTM models.

II. Related Work: ML-Based Solar Inverter Control

A. Single-Phase Solar Inverter Description

Fig. 1 illustrates a single-phase solar inverter using ML-based current loop vector control. I_d and I_q currents are computed from the inverter voltage (V_g) and inverter current

This work was supported in part by the National Science Foundation (NSF) under award No. CNS-2131163, CNS-2131070.

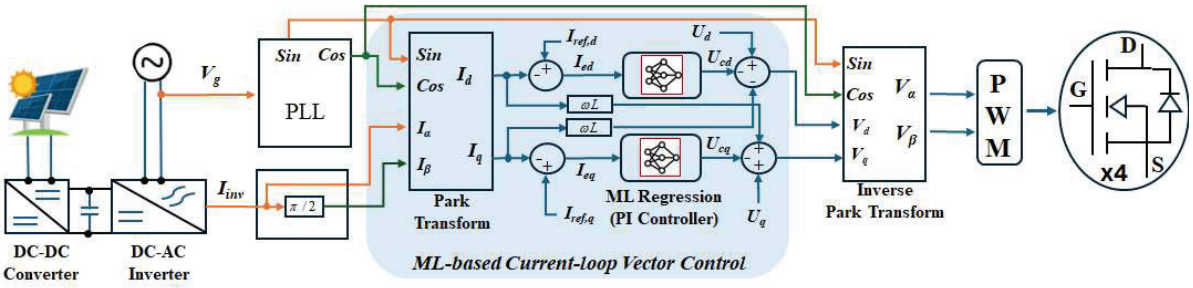

Fig. 1. ML-based decoupled vector current loop controller for a single-phase solar inverter.

(I_{inv}) through a phase locked loop (PLL) and Park Transform block. The error signals of I_d and I_q, derived from their respective reference signals, are fed into two ML regression-based models, replacing the conventional decoupled vector PI controllers. The resulting V_d and V_q signals are then transformed through an Inverse Park Transform block, with V_α serving as the duty cycle signal for PWM signal generation.

B. ML Regression-Based Model Candidates

Candidates for the ML regression models include a neural network-based regression method [4], LSTM, GRU, and Bi-LSTM which can capture the non-linearity characteristics of the controller without requiring deep knowledge of the system dynamics. For example, the LSTM regression model can predict the PI controller output Y based on input data x as follows:

$$Y = bias_y + W_y * a(bias_x + W_h * h_{t-1} + W_x * x_t) \quad (1)$$

where W_y is the weight value of the current hidden layer; a is denoted as the activation function, tanh; and W_h and W_x correspond to weight of the previous hidden layer and the weight of the current input, respectively.

III. Proposed Adversarial ML Attack Modeling

The goal of an AML attack targeting the solar inverter is to mislead the ML regressions by generating stealthy malicious input data, X_{adv}. X_{adv} can be created by injecting η into the I_{ed}. Fig. 2 shows the potential attack points of the inverter controller and corresponding tactics to inject η into the current-loop vector control and the negative impact of the X_{adv} in the controller. For example, an adversary ML attack algorithm can be injected in the new controller firmware [5] and executed to add η. Table I shows two white-box adversarial ML attack algorithms, FGSM and BIM.

In the FGSM, an adversary can choose the strength of data perturbations that is defined by ϵ. A_{ML} signifies the mean squared error (MSE) loss function between the true testing output data and the model's predictions. ∇_x corresponds to taking the gradient of the model's loss. The gradient data is then reshaped into a format that is accepted for ML testing on unseen data. The sign of the reshaped gradient data is taken and multiplied by epsilon which defines the malicious perturbations introduced into the original data that is now defined as X_{adv}. The BIM introduces three attack parameters, α, *num_iterations* and ϵ. The attack is performed at an iterative level depending on the number of iterations specified by the adversary with the amount of perturbation defined by α. In addition, ϵ now serves as a value chosen that serves to ensure the perturbations stay within a limit of the original data. This allows stealthier adversarial perturbations with the tradeoff of increased computational cost to perform the attack due to the iterative nature.

TABLE I
Two Adversarial ML Attack Generation Algorithms

Algorithm 1: FGSM adversarial ML attack

Input: Original sensor data X and its $\hat{Y}$
Output: Perturbed sensor data X_{adv}
Data: attack parameter: ϵ

$A_{ML} = (mean((Y_{test} - Y_{Predicted}).^2))$;
$gradient \leftarrow (\nabla_X A_{ML}(X, \hat{Y}))$;
$k \leftarrow reshape(repmat(gradient, [3,1,1]), [3,1,2500$
$\eta \leftarrow \epsilon \cdot sign(k)$;
$X_{adv} \leftarrow X + \eta$;
return X_{adv};

Algorithm 2: BIM adversarial ML attack

Input: Original sensor data X and its $\hat{Y}$
Output: Perturbed sensor data X_{adv}
Data: attack parameters: ϵ, α, num_iterations

for *iter = 1:num_iterations* **do**
  $A_{ML} = (mean((Y_{test} - Y_{Predicted}).^2))$;
  $gradient \leftarrow (\nabla_X A_{ML}(X, \hat{Y}))$;
  $k \leftarrow reshape(repmat(gradient, [3,1,1]), [3,1,2500$
  $\eta \leftarrow \alpha \cdot sign(k)$;
  $X_{adv} \leftarrow X + \eta$;
  $X_{adv} = \min(\max(X, X - \epsilon), X + \epsilon)$;
end
return X_{adv};

A. Malicious Firmware Appraoch

In [4], a RNN was successfully deployed onto a Texas Instrument (TI) Solar Evaluation Kit's digital signal processor (DSP) controller (TI TMS320F28335 MCU) by updating the current loop control portion of the MCU's programming which originally contained the 3 pole 3 zero (3p3z) controller and flashing the DSP's firmware using TI's Code Composer Studio (CCS) and TI's UniFlash tool. Since it is possible to flash the DSP to provide firmware updates to the control hardware, it is realizable that if an attacker gains physical access to the control hardware, they may create malicious firmware that spoofs

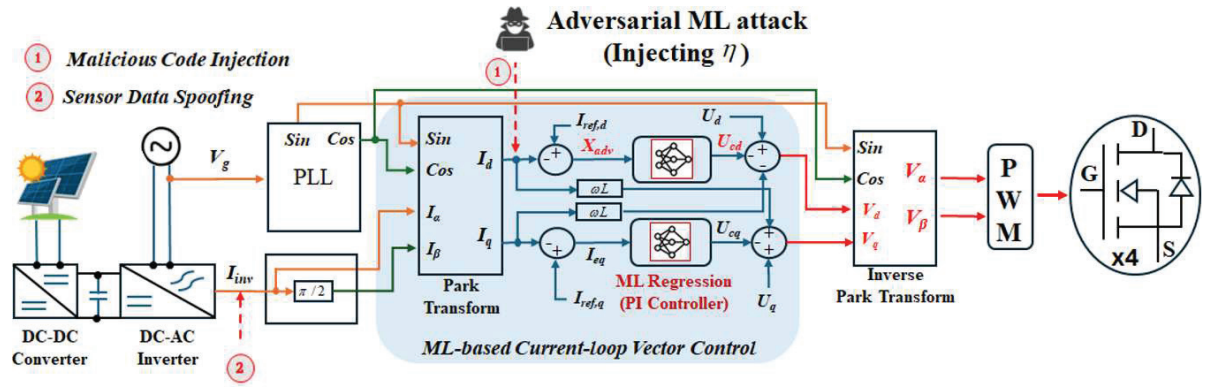

Fig. 2. Attack modeling for adversarial ML attacks.

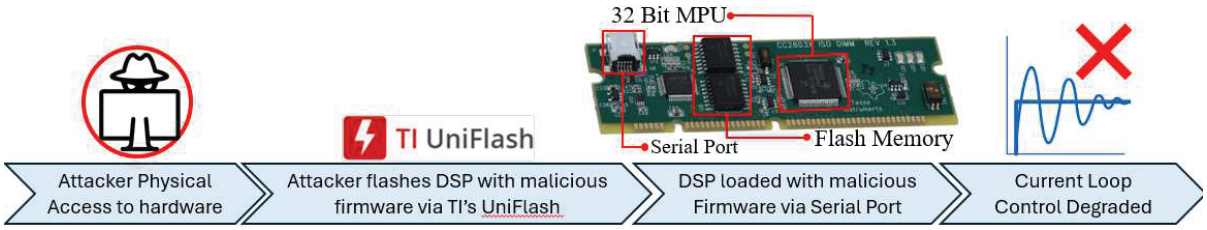

Fig. 3. Proposed Malicious Firmware Pipeline

sensor data inputs such as I_d and I_q currents to provide a malicious signal that will result in degraded current loop control. It is also realizable that an attacker may instead choose to inject η perturbations to disrupt the control defined as a function in the malicious firmware code. This can be achieved by accessing and modifying the control hardware's firmware files, and re-flashing the hardware using UniFlash to create a malicious executable that would run on the control hardware.

IV. Validation

In this paper, a single-phase solar inverter was designed in MATLAB/Simulink and three ML regression-based models (LSTM, GRU, and Bi-LSTM) for current-loop control are trained in the MATLAB Simulink environment. Then BIM-based adversarial ML attack and FGSM-based adversarial ML attack were applied to the PI controller and deployed ML-based controllers. Adversarial ML attack parameters: ϵ, α, and iterations were chosen to highlight the critical threshold of the model's predictions. FGSM utilized an ϵ of 0.6. BIM used an ϵ of 0.8, α of 1.49 and 630 iterations. The impact of the two AML attacks was compared in terms of prediction accuracy of inverter terminal voltage in the d axis, V_d with performance metrics (root mean squared error (RMSE), mean squared error (MSE), and mean absolute error (MAE)). Fig. 4 shows the comparison of performance of the controllers in normal case (Fig. 4(a)) and under BIM-based adversarial attack case (Fig. 4(b)). In normal case, all methods show similar performance, as shown in Fig. 4(a). Fig. 4(b) clearly depicts that adversarial ML attack perturbations has negligible impact observed on the PI controller while significantly disrupting ML-based controllers. Table II highlights the numerical performance among the three ML models. as well as highlighting the performance under two attacks. BIM had a stronger impact on the performance at the

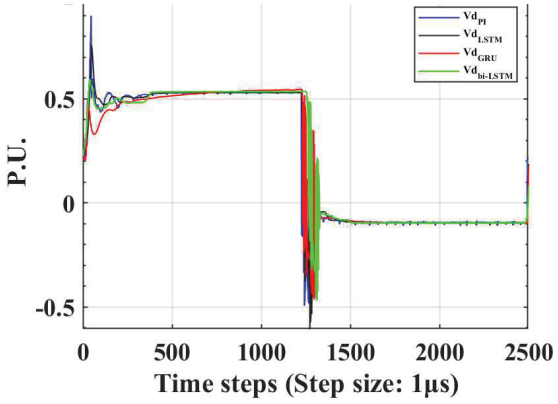

Fig. 4. Comparison of controller performance among a reference PI controller and ML controllers: (a) normal and (b) BIM attack.

TABLE II
PERFORMANCE COMPARISON OF ML MODELS

	LSTM				GRU				Bi-LSTM			
	Normal		AML Attack		Normal		AML Attack		Normal		AML Attack	
	Train	*Test*	*FGSM*	*BIM*	*Train*	*Test*	*FGSM*	*BIM*	*Train*	*Test*	*FGSM*	*BIM*
RMSE	0.0769	0.1392	0.3379	0.3897	0.0593	0.1403	0.6836	0.9516	0.0535	0.1038	0.1271	0.1222
MSE	0.0059	0.0194	0.1142	0.1519	0.0035	0.0197	0.4672	0.9056	0.0029	0.0108	0.0162	0.0149
MAE	0.0280	0.0401	0.2645	0.2587	0.0419	0.0752	0.5334	0.7001	0.0161	0.341	0.653	0.0760

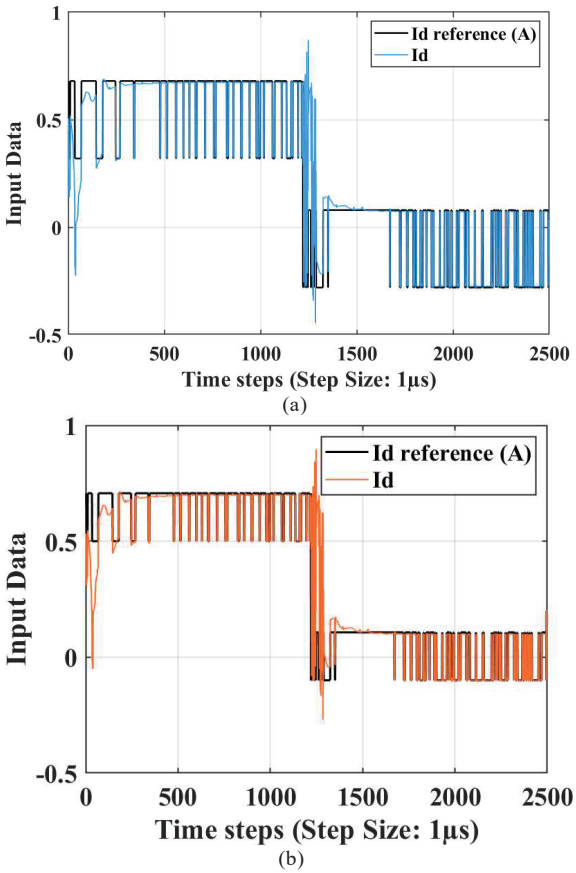

Fig. 5. Comparison of injected AML data perturbations on Id reference and Id current data: (a) FGSM and (b) BIM.

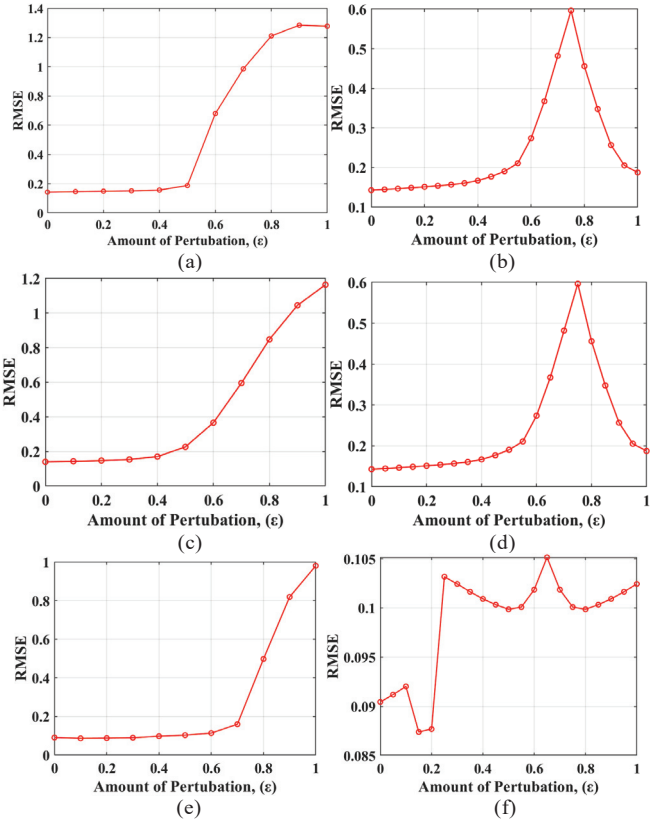

Fig. 6. Comparison of RMSE Loss versus amount of perturbation for FGSM and BIM attacks respectively: (a) LSTM-FGSM, (b) LSTM-BIM, (c) GRU-FGSM, (d) GRU-BIM, (e) Bi-LSTM-FGSM, and (f) Bi-LSTM-FGSM.

cost of computation, however the perturbated data introduced is much stealthier than the FGSM approach. Bi-LSTM regression performance shows highly accurate control under the adversarial ML attacks. It is noted that choosing/designing ML-based controller robust to potential adversarial ML attacks needs to be considered.

Fig. 5 illustrates the visual comparison of adversarial ML attack perturbations resulting from FGSM and BIM attacks respectively. Both AML algorithm's goal is to inject targeted malicious input data to each ML model. As we can observe from Fig. 5 (a) waveform, it introduces perturbations that range from above and below the original time series data being processed by the ML models. To be more specific, the input data that is relevant for each ML model to predict an output voltage in the d/q axis suitable for PWM generation consists of a current reference values defined by the inverter's software, the real time current value, as well as the error between the two signals. This input information is relevant for ML model predictions so that the model may perform control on the signals. In the simulation, a reference current is set to 0.5A in the d axis. FGSM introduces some perturbations above and below this value depending on the ε injected by the attacker, while BIM iteratively clips the perturbations to stay within a window only a bit above the reference current and real time current to improve stealthiness.

Fig. 6 highlights RMSE loss versus the amount of perturbation injected between FGSM and BIM for each model. Note that for every BIM iteration, ε is varied from 0 to 1 just as FGSM, however α was set to a constant 1.49 to highlight the highest effect on RMSE loss. It should be noted there is a

balance where the perturbation is strong enough to cause maximum disruption to the model but not so strong that the model begins to adapt or mitigate the perturbation. This results in a peak RMSE loss. For FGSM, this peak occurs because, beyond a certain point, the perturbations become excessively large and easily detectable, leading to diminishing returns in terms of increasing RMSE. For BIM, the peak is reached when the cumulative effect of iterative small perturbations reaches its most effective disruption point without becoming overly noticeable.

V. Conclusion

In conclusion, this paper discussed regression based ML models, while additionally evaluating each model's robustness to an AML attack. More HIL experiments will be investigated using FPGA hardware for real time control and validation. Utilizing DSP hardware may allow the realization of malicious firmware approach to observe the effects of control before and after the attack. Both approaches aim to demonstrate practical real hardware-based ML control system applications and highlight cybersecurity importance.

References

[1] S. Zhao, F. Blaabjerg and H. Wang, "An overview of artificial intelligence applications for power electronics," *IEEE Trans. Power Electronics*, vol. 36, no. 4, pp. 4633-4658, April 2021.

[2] S. D. Lalitha, C. Danamaraju, M. P. Raj, D. Kumar, R. Sethuraman and M. V, "AI-driven control strategies for power electronics converters," *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Chennai, India, 2024, pp. 1-6.

[3] A. De La Cruz, J. Zeng, T. Kim and V. Winstead, "Comparing support vector machine and artificial neural networks based model predictive control in power electronics," in *Proc. 2023 IEEE Energy Conversion Congress and Exposition (ECCE)*, Nashville, TN, USA, 2023, pp. 3490-3494.

[4] C. Hingu, X. Fu, S. Smith, T. Saliyu and L. Qingge, "FPGA acceleration of a real-time neural network controller for solar inverter," in *Proc. 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, USA, 2022, pp. 0413-0420.

[5] R. Huang and Y. Li, "Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system," *IEEE Trans. Smart Grid*, vol. 14, no. 3, pp. 2367-2376, May 2023.

[6] J. Qi, A. Hahn, X. Lu, J. Wang, and C.-C. Liu, "Cybersecurity for distributed energy resources and smart inverters," *IET Cyber-Phys. Syst., Theory Appl.*, vol. 1, no. 1, pp. 28–39, Dec. 2016.

[7] B. Ahn, A. M. Jenkins, T. Kim, J. Zeng, L. McLauchlan and S-W. Park, "Exploring ransomware attacks on smart inverters," in *Proc. 2023 IEEE Energy Conversion Congress and Exposition (ECCE)*, Nashville, TN, USA, 2023, pp. 1567-1573

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997.